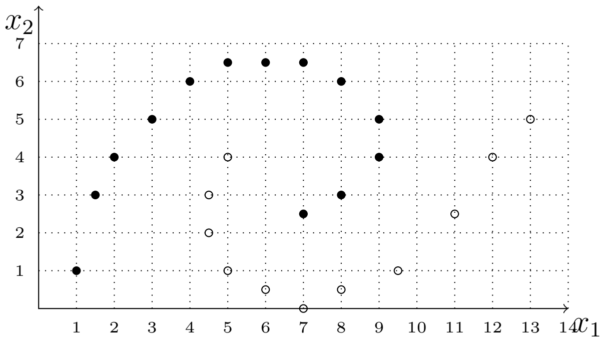

Figure 1: Two entangled curves

Homework 5

posted April 27, 2020
due May 13, 2020

1. Let $X = \{-10, -9, -7, -6, -1, 0, 2, 6, 7, 9\}$ be a set of 10 numbers in $\mathbb{R}$. Draw the singke-link dendrogram for this set.

2. Let $S = \{x_1, \ldots, x_n\}$ be a set of n points in $\mathbb{R}$. Prove that the largest height of a dendrogram constructed for S is $n - 1$ and the smallest is $\lceil \log_2 n \rceil$.

3. Consider the set of points that consists of two "entangled spirals" shown in Figure 1.

   Extract the coordinates of the points from Figure 1 and compute the `dist` object using the Euclidean metric. Show that the sets of points of the two curves can be separated by using the single link method; in other words, it is possible to cut the dendrogram obtained by this method such that the points of the two curves belong to two different clusters. Verify that this is not possible if we use the complete link and explain why is this the case.

4. If `m` is a matrix of dissimilarities between objects, then `m` is a symmetric matrix that has 0s as its diagonal elements. To create a `dist` object as required by the `hclust` function we can use the coercion `as.dist` by writing `d <- as.dist(m)`.

Start from the matrix

$$m = \begin{pmatrix} 0 & 1 & 5 & 8 & 3 & 4 \\ 1 & 0 & 2 & 2 & 8 & 6 \\ 5 & 2 & 0 & 1 & 7 & 9 \\ 8 & 2 & 1 & 0 & 4 & 2 \\ 3 & 8 & 7 & 4 & 0 & 7 \\ 4 & 6 & 9 & 2 & 7 & 0 \end{pmatrix}$$

Starting from the matrix `m` given above construct a single-link clustering and a complete-link clustering. Note that the dendrograms of these clusterings are distinct. What is the least number of entries of `m` that you need to change to obtain the same dendrogram using these two distinct methods?

5. Prove that the k-means algorithm applied to a set of n points may require no more than $O(k^n)$ iterations.